Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 7(C) (2010) 406–413

Procedia Social and Behavioral Sciences

International Conference on Learner Diversity 2010

Detecting Gender Related DIF using Logistic Regression and Mantel-Haenszel Approaches

Nabeel Abedalaziz*

Faculty of Education, University of Malaya,50603 Kuala Lumpur, Malaysia

Abstract

The study was conducted to find out the agreement between two approaches (i.e. logistic Regression model, and M-H) in detecting a gender-related differential item functioning of a mathematical ability scale items. The scale was developed and administered to samples of 800 males and females students (380 males and 420 females) in Jordan. The study pointed out: (1) the percentage of agreement between the two approaches in detecting DIF was 80%. (2) Males outperformed females in spatial and deductive abilities, whereas females outperformed males in numerical ability.

Keywords: Mantel-Haenszel; Area index; Mathematical ability; One-parameter logistic model

1. Introduction

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1990). When personal attributes, such as gender systematically affect examinee achievement on an item, the result can be differential item functioning (DIF).

1.1. Differential Item Functioning

One of the important issues faced by counseling psychologists is that of responding to the diversity of clients. In particular, it is important that the tests used by counseling psychologists be free of systematic demographic subgroup bias. Item Response Theory techniques provide a powerful means of testing items for bias, using what is known as differential item functioning (DIF) analysis. In contrast, Classical Test Theory based methods of assessing bias are fundamentally limited, especially approaches that base their assessment of bias on the presence of group mean differences in total tests scores across demographic groups. In essence, such methods cannot distinguish between the situation in which (a) the subgroups have different means, and the test is biased, versus (b) the means differ, but the test is not biased.

Two different forms of DIF have been recognized. These have been called uniform and non-uniform DIF. Uniform DIF is said to apply when differences between groups in item responses are found at all ability levels, while in non-uniform DIF an interaction is found between ability level, group assignment, and item responses (Harvey & Greenberg, 1996; Sells, 1973). Simulation studies from educational testing experts have found that LR-based DIF detection techniques enables the detection of both uniform and non-uniform DIF, while Mantel-Haenszel techniques are better suited for the analysis of uniform DIF (Sells, 1973; Anogoff & Ford, 1973). Several approaches have been promulgated for the statistical assessment of DIF. Most approaches or DIF were developed in educational settings in which items are generally dichotomously scored as correct or incorrect. Mantel–Haenszel (MH)- based techniques were initially applied to the problem of assessing DIF.

*Corresponding author.Tel.: +603-8921-6266; fax:+603-8925-4372
E-mail address: nabel@um.my

1877-0428 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2010.10.055

It was recognized by the early 1990s that logistic Regression (LR) based techniques were more powerful than MH-based techniques (Sells, 1973; Anogoff & Ford, 1973).

1.2. Gender differences in mathematical ability

One line of research focused on the relationship between three cognitive abilities (verbal, quantitative, and visual-spatial abilities) and gender differences in mathematical ability. However, evidence from these studies is inconsistent and sometimes conflicting.

Spatial abilities generally refer to skill in representing, transforming, generating and recalling symbolic, nonlinguistic information" (Linn and Petersen, 1985). Spatial skills involve the ability to think and reason using mental pictures rather than words (Nuttall, Casey, and Pezaris, 2005). They are believed as one important component of mathematical thought during mathematical problem solving (Battista, 1990; Casey, 2003; Halpern, 2000).

Spatial visualization has been defined as "those spatial tasks which involve complicated multistep manipulations of spatially presented information" (Linn and Petersen, 1985). Although many researchers have found that spatial visualization and problem-solving were related (Battista, 1990; Fennema and Tartre, 1985; Sherman, 1979). Studies investigating gender differences in spatial visualization have reported inconsistent results. Ben- Chaim et al. (1988) found that there were statistically significant gender differences in spatial visualization among middle school students; while other researchers concluded that gender differences in spatial visualization were small or null among middle school students (Armstrong, 1980; Fennema and Sherman, 1977, 1978; Linn and Petersen, 1985; Tartre and Fennema, 1995; Voyer et al., 1995).

Mental rotation refers to the ability to transform mentally and manipulate images when the object is rotated in three-dimensional space (Nuttall et al., 2005). Many studies suggested that there was a large gender difference in mental rotation ability with males outperforming females (Casey et al., 1995; Halpern, 2000; Linn and Petersen, 1985; Masters and Sanders, 1993; Voyer, et al., 1995). Evidence from a variety of sources has shown that there were gender differences in verbal skills with females outperforming males on many verbal tasks (Maccoby and Jacklin, 1974; Halpern, 2000). However, Hyde and Linn (1988) concluded that gender differences in verbal abilities had declined and were negligible now.

Studies that reported gender differences in mathematical abilities favoring males had generally consistent conclusions. Linn and Hyde (1989) concluded that females are superior at computation at all ages and that differences favoring males on problem solving emerge in high school. Benbow and Stanley (1980) indicated that gender difference in mathematical reasoning ability in favor of boys was observed before girls and boys started to differ in mathematics courses taking. This gender difference even increased through the high school years. Benbow and Stanley (1983) also suggested that males dominated the highest end on mathematical reasoning ability before they entered adolescence. Some studies have attributed gender differences in quantitative SAT achievement to males and females' differential patterns of course taking. They suggested that increasing female's high-level mathematical course-taking would effectively increase their achievement in quantitative SAT.

Students taking higher level mathematics courses would benefit from training in abstract reasoning, from computational practice, and from generally being more comfortable in working with numbers. (Pallas and Alexander, 1983). This explanation was in conflict with the conclusion of Benbow and Stanley (1980), who found that gender difference in mathematical reasoning ability in favor of boys, was observed among gifted youth before they started to differ in mathematics courses taking. The inconsistent conclusion might be due to the different samples they used.

Methods for detecting item bias have proliferated in recent years and have been reviewed by Petersen (1977), and Rudner (1977). The various methods include techniques that examine (a) differences in relative item difficulty across different groups, (b) differences in item discrimination across groups, (c) differences in the item-characteristic curves for different groups, (d) differences in the distribution of incorrect responses for various groups, (e) differences in multivariate factor structures across groups (Subkoviak et al ,1987). Thus, the researcher wishing to select a bias detection method is confronted with many methods and no clear guidelines for choosing among them. The comparison of bias methods is an important practical concern. Rudner (1977) and Scheuneman (1979) have noted the need to empirically compare the various methods.

1.3. Significance of the study

The significance of this study is related to the importance of mathematical ability in the current mathematics education reform and the goal of achieving equal educational outcomes in students' learning of mathematics (National Council of Teachers of Mathematics (NCTM), 1989, 2000). Since mathematics is no longer just a prerequisite subject for prospective scientists and engineers but is a fundamental aspect of literacy for the twenty-first century (Mathematics Sciences Education Board, 1993; NCTM, 1989), male and female students should have equal opportunity to learn mathematics, have equal treatment within classrooms, and achieve equal mathematics educational outcomes (Fennema & Leder, 1990). The uniqueness of this study was its investigation of gender differences of relatively large samples of Jordanian students using mathematical ability test.

The present study sought answers to the following questions: (1) to what extent do the two methods (i.e. logistic regression model and Mantel-Haenszel) agree or disagree in the identification DIF? (2) Are there gender differences in mathematical ability? Are gender differences linked to content areas within mathematics?

2. Method

2.1. Participants

A total of 800 (380 males and 420 females) tenth grade students in Jordan were targeted as participants in this study, during the ending period of the First semester, school year 2009- 2010.

2.2. *Instrument*

A mathematical ability scale was developed as a part of this study. The scale compressed of 30 multiple-choice items to measure three components of mathematical ability (i.e. numerical ability, deductive ability, and spatial ability). Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be 0.91. Spearman-Brown Correction on split-half reliability for odd even comparison also show similar results r = 0.89. Validity of the instrument was shown using inter-correlation of the scale (0.19 to .855). Factor Analysis reveals that the test measure one trait (unidimensionality).

2.3. *Detecting DIF*

In the present study, two techniques have been used (i.e. logistic regression, and Mantel-Haenszel).

2.3.1. *Logistic Regression (LR)*

Swaminathan and Rogers (1990) applied the Logistic Regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and nonuniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and nonuniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of nonuniform DIF compared to the Mantel-Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). These studies also identified two major weaknesses in the LR DIF procedure: 1) the Type I error or false positive rate was higher than expected, and 2) the lack of an effect size measure.

Logistic regression has a formal mathematical equivalence to the log linear model approach of Mellenbergh (1982): Coefficients for group, total score, and interaction terms are estimated and tested for significance with a model comparison strategy. However, logistic regression is highly similar to standard ordinary least squares regression. It can be conceptualized as an equation that uses group, ability, and group-by-ability terms to predict whether an item response is right (1) or wrong (0). This property is desirable for didactic purposes.

Logistic regression uses the examinee as the unit of analysis, and has the following form:

$$P(u/x,g) = \frac{e^{[1-u][-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}{1+e^{[-\beta_0-\beta_1 x-\beta_2 g-\beta_3(xg)]}}$$

Where:

g: represents group membership (0 for focal group (female) and 1 for reference group (male)).
x: the matching group (the observed total test score).
u: represents the item response value (0 for an incorrect answer and 1 for correct answer).
xg: represents the interaction between the matching variable and the group variable..
β_0: $\beta_1\beta_2$ and β_3: Parameters to be estimated.

The above equation is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership. Once the estimates of the four coefficient parameters, β_0: $\beta_1\beta_2$ and β_3, for an item are obtained from a sample of test responses, the usual likelihood ratio chi-square tests of significance of the estimates of β_2 and β_3 are conducted to examine if DIF exist. The null hypothesis is that $\beta_2 = \beta_3 = 0$. An item shows uniform DIF if $\beta_2 \neq 0$ and $\beta_3 = 0$ with 1 degree of freedom and nonuniform DIF if $\beta_3 \neq 0$ (whether or not $\beta_2 = 0$) with 1 degree of freedom (Swaminathan & Rogers, 1990).

In the present study, the item reveals uniform DIF when the significant odd ratio is for the group, whereas the item reveals nonuniform DIF when the significant odd ratio is for the interaction between the group and total score. The item reveals DIF in favor of males when the significant odd ratio is greater than one, whereas the item reveals DIF in favor of females when the significant odd ratio is less than one (α=0.05).

2.3.2. Mantel-*Haenszel method (M-H)*

The Mantel-Haenszel (M-H) procedure was originally used to match subjects retrospectively on cancer risk factors in order to study current cancer rates (Mantel & Haenszel, 1959). The procedure has since been adapted to study differential item functioning and is now the primary DIF detection device used at the Educational Testing Service (ETS; Dorans & Holland, 1992). The M-H method works by first dividing subgroups into the reference group (e.g., males) and the focal group (e.g., females). The focal group is of primary interest in the analysis and is compared to the reference group after being matched on θ (Uttaro & Millsap, 1994). The total test score usually serves as the θ estimate, and the performance (i.e. item endorsement rates) of the reference and focal groups is compared at unit intervals of θ weighted by the number of examinees at each level (Scheuneman & Gerritz, 1990). From this comparison, an odds-ratio estimator can be calculated, and a χ^2 test of significance can be carried out to assess the presence of DIF.

To assess the degree of DIF present, the odds-ratio estimator can be transformed onto the ETS "delta metric" (D; Dorans & Holland, 1992). The D statistic represents the difference in item difficulty for the reference and focal groups after the total score has been taken into account (Scheuneman & Gerritz, 1990). The advantage of using the D statistic to classify degree of DIF

present is that the ETS has defined the values of it into a classification scheme delineated by Dorans and Holland (1992). A D value of 0.0 indicates no DIF, a positive value indicates DIF favoring the focal group (e.g., females), and a negative D value reflects DIF that favors the reference group (e.g., males). More specifically, there are three possible degrees of DIF: (a) negligible DIF, where χ^2 is nonsignificant or the absolute value of D is less than 1.0; (b) intermediate DIF, where χ^2 is significant and D is between 1.0 and 1.49 in absolute value; and (c) large DIF, where χ^2 is significant and the absolute value of D is 1.5 or larger (Dorans & Holland, 1992).

The Mantel-Haenszel technique is ideal because it does not rely solely on the χ^2 statistic, which can be overly sensitive when large samples are used, which is customary in DIF analyses. The D statistic not only complements the χ^2 statistic, but also allows assessments of the degree of DIF to be made. The one limitation of the M-H procedure is that it may lack power to detect DIF that is not uniform across the range of θ scores (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994).

3. Results

Table 1 shows the summary results of the Logistic Regression method to identify Differential Item Functioning on the mathematical ability scale for each of the thirty items. Eighteen of thirty items or 57 percent of the items revealed DIF (i.e. the items: 1, 2, 4, 13, 14, 19, 24, 27, 29, and 30 were revealed uniform DIF, whereas the item: 3, 6, 8, 9, 15, 20, 25, and 26 were revealed nonuniform DIF). The items: 1,2,3,6,8,19, 24, 27, 29, and 30 were in favor of males, whereas the items: 9, 13, 14, 15, 20, 24, 25, and 26 were in favor of females).

Table 1. Summary results from the logistic regression method to identify differential item functioning on the mathematics ability scale

Item	Variable	Statistical significance	Odds-Ratio	Type of DIF
1	Group	0.01	2.69	Uniform
	Interaction	0.11	0.94	
2	Group	0.05	1.26	Uniform
	Interaction	0.13	0.56	
3	Group	0.11	0.91	Nonuniform
	Interaction	0.01	2.68	
4	Group	0.14	0.51	
	Interaction	0.70	0.98	
5	Group	0.26	1.68	
	Interaction	0.88	0.98	
6	Group	0.01	2.70	Nonuniform
	Interaction	0.11	0.94	
7	Group	0.62	1.22	
	Interaction	0.78	1.02	
8	Group	0.01	1.77	Uniform
	Interaction	0.54	0.99	
9	Group	0.53	1.30	
	Interaction	0.00	0.86	Nonuniform
10	Group	0.26	0.78	
	Interaction	0.88	0.99	
11	Group	0.99	0.98	
	Interaction	0.07	1.13	
12	Group	0.70	0.85	
	Interaction	0.58	1.03	
13	Group	0.02	0.12	Uniform
	Interaction	0.13	1.05	
14	Group	0.03	0.33	Uniform
	Interaction	0.12	1.04	
15	Group	0.07	3.52	Nonuniform
	Interaction	0.02	0.88	
16	Group	0.60	1,45	
	Interaction	0.66	0.97	
17	Group	0.11	0.28	
	Interaction	0.22	1.09	
18	Group	0.66	0.83	
	Interaction	0.38	1.05	
19	Group	0.03	2.69	Uniform

	Interaction	0.14	0.92	
20	Group	0.53	1.30	Nonuniform
	Interaction	0.00	0.86	
21	Group	0.26	1.68	
	Interaction	0.88	0.90	
22	Group	0.63	1.23	
	Interaction	0.78	1.01	
23	Group	0.53	0.74	
	Interaction	0.73	0.98	
24	Group	0.00	0.86	Uniform
	Interaction	0.19	0.72	
25	Group	0.16	0.04	Nonuniform
	Interaction	0.00	1.23	
26	Group	0.07	3.52	
	Interaction	0.02	0.88	Nonuniform
27	Group	0.04	0.78	Uniform
	Interaction	0.19	0.82	
28	Group	0.03	0.92	
	Interaction	0.21	1.62	
29	Group	0.02	2.72	Uniform
	Interaction	0.42	0.82	
30	Group	0.04	1.62	Uniform
	Interaction	0.23	0.92	

Table 2 shows the summary results from the Mantel-Haenszel method to identify Differential Item Functioning on the Mathematics Ability Scale. The M-H procedure flagged sixteen of thirty or 53 percent of the thirty items as indicating DIF (the items: 13, 14, 24, 25, 27, and 28 were in favor of females and the items: 1, 2, 3, 6, 8, 18, 19, 26, 29, and 30 were in favor of males).

Table 2. Summary results from the mantel-haenszel method to identify differential item functioning on the mathematics ability scale

	Group mean				
Item	Male	Female	Δ	Chi-square	p-value
1.	0.62	0.55	-1.05*	77.69	0.00
2.	0.61	0.56	-1.04*	77.68	0.00
3.	0.82	0.77	-1.15*	74.36	0.00
4.	0.70	0.78	0.72	29.43	0.00
5.	0.61	0.55	-0.98	75.81	0.00
6.	0.81	0.76	-1.12*	73.36	0.00
7.	0.70	0.72	0.21	2.62	0.11
8.	0.86	0.83	-1.11*	0.39	0.53
9.	0.41	0.40	-0.84	37.30	0.00
10.	0.48	0.50	0.21	3.29	0.07
11.	0.76	0.78	0.42	7.39	0.01
12.	0.61	0.71	0.84	42.63	0.00
13.	0.62	0.75	1.65*	142.45	0.00
14.	0.63	0.73	1.96*	123.21	0.00
15.	0.38	0.44	0.40	6.76	0.01
16.	0.75	0.81	0.51	10.05	0.00
17.	0.27	0.40	0.82	44.69	0.00
18.	0.18	0.12	-1.96*	123.21	0.00
19.	0.70	0.64	-1.11*	90.31	0.00
20.	0.85	0.89	0.66	15.05	0.60
21.	0.57	0.61	0.07	0.28	0.00
22.	0.50	0.50	-0.49	17.54	0.00
23.	0.53	0.57	0.02	0.01	0.87
24.	0.57	0.78	1.03*	51.77	0.00
25.	0.58	0.66	1.68*	198.95	0.00
26.	0.32	0.22	-1.02*	57.78	0.00
27.	0.43	0.56	1.76*	248.17	0.00
28.	0.31	0.35	1.21*	90.81	0.00
29.	0.84	0.73	-1.35*	114.68	0.00

30.	0.58	0.40	-1.54*	173.45	0.00

*The item reveals DIF

In order to inspect the consistency between the two approaches in detecting DIF for the test, the percentage agreements between the two approaches were computed (i.e. the degree of correspondence between the two approaches with respect to the items revealing or not revealing the DIF for all items were computed).

Table 3 summarizes the consistency in which the M-H and Logistic regression methods flagged the items. The two methods were agreeable in allocating fourteen items as revealing DIF, and ten items as not revealing DIF. As such, the percentage of agreement between M-H and Logistic regression methods is 80% (i.e. 14+10/30=80%).

Table 3. Pair wise agreement between m-h and logistic regression methods.

	Results From Mantel-Haenszel		
Results From Logistic regression	No. of Nonflagged Items	No. of flagged Items	Marginal Total
No. of nonflagged items	10	2	12
No. of flagged items	4	14	18
Marginal total	14	16	30

3. Discussion and conclusion

In summary, the percentage of agreement between the two approaches in detecting DIF are relatively high, however, this may due to: both methods related to classical theory of measurement. This finding seems to be consistent with the previous studies (e.g.: Intaswan, 1979; Seong & Subkoviak, 1987; Hambleton &Rogers, 1989; Baghi & Ferrara, 1989; Skaggs & Lists, 1992; Hakim & Cohen, 1995; Stage, 2000).

The DIF indices point to the conclusion that females had an advantage over males on the numerical ability, whereas males had an advantage on items involving spatial ability and deductive ability, The tendency for males to perform better than females on spatial ability and inductive ability, and women to perform better on numerical ability is consistent with previous findings (e.g. Willson, Fernandez and Hadaway, 1993; Gallagher, DeLisi, Holst, McGillicuddy-DeLisi, Morely and Cahalan, 2000) .

In previous studies, however, females usually performed better on Number and Computation. The fact that this test was tied to a specific curriculum did not appear to help females' performance The Researchers consistently found that male students are superior in geometry and visualization (Geary, 1996).On the other hand, female show superiority in computation based on the data available. Gender differences in achievement in mathematics in favor of boys have been found in standardized tests and are most prominent at the very high levels of achievement (Leder, 1992). These differences are likely to both content and ability dependent. While males outperform females in scientific and mathematical tasks, females outperform males in tasks involving verbal abilities.

There are many studies that focus on differences between men and women in tests (Gallaghe, De Lisi, Holset, Mc Gillicuddy-De Lisi, Morly & Cahalan, 2000; Kimball, 1994; Willingham & Cole, 1997). From the findings of the present study and the earlier studies, one conclusion can be drawn is that males have a better spatial ability than females (Geary, 1996). Males use this spatial more often than females when solving problems, which can give advantages while solving certain kinds of problems in geometry (Geary, 1996). Many studies indicate that women are better than men in verbal skills, which can give them advantages on items where communication is important. Women also score relatively higher on tests in mathematics that better match coursework. Men tend to outperform women in geometry and in arithmetic and algebraic reasoning questions. Women tend to be better at intermediate algebra and arithmetic and algebraic operations (Willingham & Cole, 1997). Gallagher, De Lisu, Holset, Mc Gillicuddy-De Lisi, Morly & Cahalan, (2000) found men outperformed women in all kind of problems, but that the differences were greater for problems requiring spatial skills or multiple solution paths than for problems requiring verbal skills or containing classroom-based content.

Spatial abilities were reported to have relationship with mathematics test scores (Casey, Nuttall, Pezaris and Benbow, 1995; Geary, Saults, Liu, and Hoard, 2000; in Nuttall et al, 2005). This relationship indicates that gender differences in spatial abilities may contribute to gender differences in mathematical problem solving.

The study provides evidence that there are gender differences in performance on test items in mathematics that vary according to content even when content is closely tied to curriculum.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Armstrong, J. M. (1980).*Achievement and Participation of Women in Mathematics: An Overview*. Denver Co: Education Commission of the States.

Baghi, Hand Ferrara, S. F (1989) A Comparison of IRT, Delate Plot, and Mantel-Haenszel techniques for detecting DIF Across Subpopulation in MTCS. ED324364.

Battista, M. T. (1990) Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education,* 21(1), 47-60.

Benbow, C. P. and Stanley, J. C. (1980) Sex differences in mathematical ability: Facts or Artifact? *Science*, 210(12), 1262-1264.

Benbow, C. P. and Stanley, J. C. (1983) Sex differences in mathematical reasoning ability: More facts. *Science,* 222, 1029-1031.

Ben-Chaim, D., Lappen, G. and Houang, R. T. (1988) The effect of instruction on spatial visualization skills of middle school boys and girls. *American Educational Research Journal*, 25(1), 51-71.

Casey, B. (2003) Mathematics problem-solving adventures: A language-arts based supplementary series for early childhood that focuses on spatial sense. In D. Clement, J. Sarama and M. A. Dibaise (eds.) *Engaging Young Children in Mathematics Education.* Mahwah, NJ: Lawrence Erlbaum Associates.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995) The influence of spatial ability on gender differences in math college entrance test scores across diverse samples. *Developmental Psychology*, 31, 697-705.

Dorans, N.J., & Holland, P.W. (1992). DIF detection and description: Mantel- Haenszel and standardization. (RR-92-10). Princeton, NJ: Educational Testing Service.

Fennema, E. & Leder, G. C. (Eds.) (1990). *Mathematics and gender.* New York: Teachers College Press.

Fennema, E. and Sherman J. (1977) Sex-related difference in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, 14(1), 51-71.

Fennema, E. and Sherman J. (1978) Sex-related difference in mathematics achievement and related factors: A further study. *Journal for Research in Mathematics Educational*, 9, 189-203.

Fennema, E. and Tartre, L. A. (1985) The use of spatial visualization in mathematics by girls and boys.*Journal for Research in Mathematics Education.* 16(3), 184-206.

Gallagher, A.M., De lisi, R., Holst, P.C., McGillicuddy-De Lisi, A.V., Morely, M. & Cahalan, C. (2000). Gender Differences in Advanced Problem Solving. Journal of Experimental Child Psychology, 75, 165-190'

Geary, D., Saults, S. J., Liu, F. and Hoard, M. K. (2000) Sex differences in spatial cognition, computational fluency and arithmetical reasoning. *Journal of Experimental Child Psychology,* 77, 337-353.

Geary, D.C. (1996). Sexual selection and sex differences in mathematical abilities. Behavioral and Brain Science, 19, 229-284.

Ha-Kim, S.,& Cohen, A. (1995). Comparison of Lord Chi Square and Rajus Measures and the Likeli hood Method on detecting of differential item functioning, Applied Measurement In education Vo14, pp (291-312).

Halpern, D. F. (2000) *Sex Differences in Cognitive Abilities* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., & Rogers, H.J. (1989). Detection potentially biased test items: comparison of IRT areas and Mantel-Haenzel methods. Applied Measurement in Education, 2, 313-334.

Harvey, R.J., & Greenberg, S.E. (1996). Gender-based differential item functioning in the Myers-Briggs Type Indicator: Implications for employee selection and big-five inventories. Unpublished manuscript. Virginia Polytechnic Institute and State University.

Hyde, J. S. and Linn, M. C. (1988) Gender difference in verbal ability. A meta-analysis. *Psychological Bulletin,* 104, 53-69.

İntassuwan, P. (1979). Comparison of three approaches for determining item bias in cross – national testing, (un published Dissertation, university of Pittsburgh. USA).

Kimball,a M.M. (1994). (It is only a myth that girls are poorer in mathematics.). Kvinnovetenskaplig tidskrift, 15(4), 39-53.

Leder, G.C. (1992). Mathematics and gender: changing perspectives. In D,A. Grouws (Ed), Handbook of Research on Mathematics Teaching and Learning. New York: Macmillan

Linn, M. C. and Hyde, J. S. (1989) Gender, mathematics, and science. *Educational Researcher,* 18(8), 17- 19, 22-27.

Linn, M. C. and Petersen, A. C. (1985) Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development,* 56, 1479–1498.

Maccoby, E. E. and Jacklin, C. N. (1974) *The Psychology of Sex Differences.* Stanford, CA: Standford University Press.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719- 748.

Masters, M. S., & Sanders, B. (1993) Is the gender difference in mental rotation disappearing? *Behavior Genetics*, 23, 337-341.

Mathematics Sciences Education Board (1993). *A conceptual guide for mathematics assessment.* Washington, DC: National Academy Press.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

National Council of Teacher of Mathematics (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Nuttall, R. L., Casey, M. B. and Pezaris, E. (2005) Spatial ability as a mediator of gender differences on mathematics test: A biological-environmental framework. In A. M.

Pallas, A. M. and Alexander, K. L. (1983) Sex differences in quantitative SAT performance: New evidence on the differential course work hypothesis. *American Educational Research Journal*, 20(2), 165- 182.

Petersen, N. S. (1977, June). *Bias in the selection rule: Bias in the test*. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Raju, N.S. (1988).The area between two item characteristic curves. Psychometrica,12, (6).

Rogers HJ, Swaminathan H. (1993) A comparison of the logistic regression and Mantel–Haenszel procedures for detecting dierential item functioning. Applied Psychological Measurement , 17(2):105–116.

Rudner, L. M. (1977). Biased Item Detecting Technique. *Journal of Educational Statistics*,5, 213-233.

Scheuneman, J.D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*, 143-152.

Scheuneman, J.D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27, 109-131.

Sells, L. W. (1973). High school mathematics as the critical filter in the job market. In R. T. Thomas (Ed.), *Developing opportunities for minorities in graduate education* (pp. 37-39). Berkeley, CA: University of California Press.

Seong, T. J, & Subkoviak, M. J. (1987). Acomparative study of recently proposed item bias detecting methods. (ERIC Document Reproducation No. E D: (365091).

Sherman, J. (1979) Predicting mathematics performance in high school girls and boys. *Journal of Educational Psychology*, 71, 242-249.

Skaggs, G., & Lissits, R. (1992). The consistency of Detecting item Bias across different test Administration: Implications of anther Failure. Journal of Educational Measurement. Vol 29 No .3.

Stage, Christina. (2000). Predicting Gender Differences in word item. A comparison of item response theory and classical test theory. A study of the swe SAT Subtest. ERIC. (Educational measurement, No 30).

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Subkoviak, M. j. Mack, J. S. Ironson, G. H. and Crag, R. D. (1987). Empirical Comparison of Selected Item bias Detection procedures with bias Manipulation. *Journal of education Measurement*, 21(1),209-223.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Tartre, L. A. and Fennema, E. (1995) Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables [Grade 6-12]. *Educational Studies in Mathematics,* 28, 199-217.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18, 15-25.

Voyer, D., Voyer, S. and Bryden, M. P. (1995) Magnitude of sex differences in spatial abilities: A metaanalysis and consideration of critical variables. *Psychological Bulletin,* 117, 250–270.

Willingham, W.W.& Cole, N.S. (1997). Gender and Fair Assessment. Lawrence Erlbaum Associates, Publishers.

Willson, J. W., Fernandez, M. L. and Hadaway, N. (1993) Mathematical problem solving. In P. S. Wilson (ed.) *Research Ideas for the Classroom: High School Mathematics.* New York: MacMillan.